

PROTECTING WIKIPEDIA FROM DISINFORMATION: DETECTING MALICIOUS EDITORS AND PAGES TO PROTECT

Francesca Spezzano
francescaspezzano@boisestate.edu

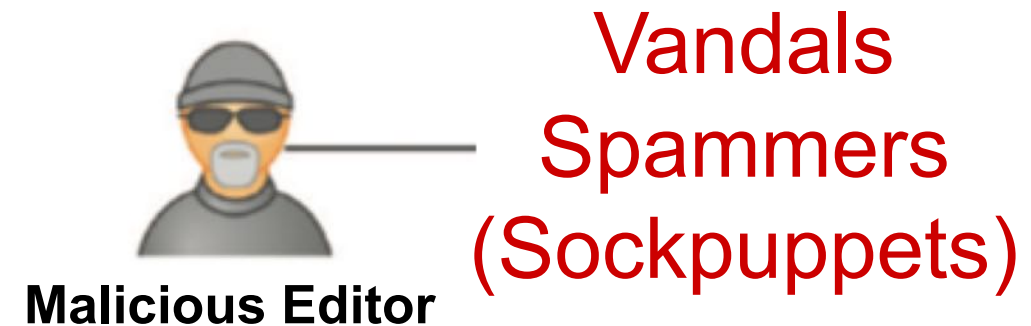
Wikimedia Research Showcase
October 16, 2019



BOISE STATE UNIVERSITY

DISINFORMATION IN WIKIPEDIA

- Freely accessible
- Large reach
- Major source of information for many

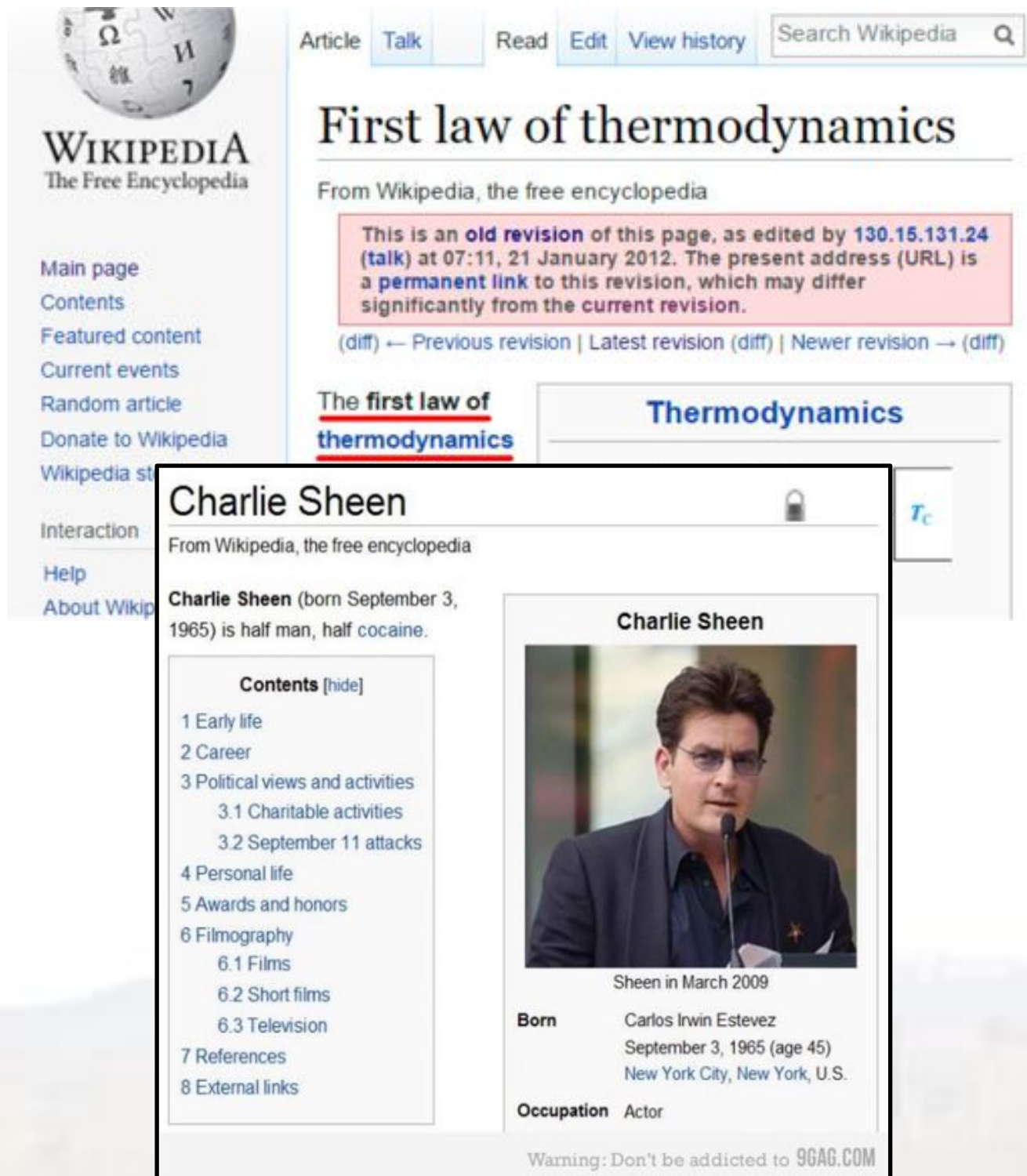


Easy to compromise article quality



The **free** encyclopedia that **anyone** can edit

DISINFORMATION IN WIKIPEDIA



The screenshot shows the Wikipedia interface. At the top, there's a navigation bar with 'Article', 'Talk', 'Read', 'Edit', and 'View history' buttons, along with a search box. The main heading is 'First law of thermodynamics'. Below it, a red box contains a message: 'This is an **old revision** of this page, as edited by 130.15.131.24 (talk) at 07:11, 21 January 2012. The present address (URL) is a **permanent link** to this revision, which may differ significantly from the current revision.' Below this, there are links for '(diff) ← Previous revision | Latest revision (diff) | Newer revision → (diff)'. The article title 'The first law of thermodynamics' is underlined in red. To the right, there's a box titled 'Thermodynamics'. Below the article title, there's a section for 'Charlie Sheen'. The text reads: 'Charlie Sheen (born September 3, 1965) is half man, half cocaine.' To the right of this text is a photo of Charlie Sheen. Below the photo, it says 'Sheen in March 2009'. To the left of the photo, there's a table of contents with links to '1 Early life', '2 Career', '3 Political views and activities', '4 Personal life', '5 Awards and honors', '6 Filmography', '7 References', and '8 External links'. At the bottom of the page, there's a warning: 'Warning: Don't be addicted to 9GAG.COM'.

Vandalism

“the act of editing the project in a malicious manner that is intentionally disruptive. Vandalism includes the addition, removal, or modification of the text or other material that is either humorous, nonsensical, a hoax, or that is an offensive, humiliating, or otherwise degrading nature.”

DISINFORMATION IN WIKIPEDIA



Spam

Unsolicited promotion of some entity

“There are three main types of spam on Wikipedia. These are:

- advertisements masquerading as articles and contributions to articles;*
- external link spamming; and*
- adding references with the aim of promoting the author or the work being referenced.”*

North Face Edited Wikipedia's Photos. Wikipedia Wasn't Happy.



In a video ad, the North Face described how it took photos of its clothing and equipment at famous outdoor destinations and uploaded the pictures to the Wikipedia pages for those locations.

DISINFORMATION IN WIKIPEDIA

Olimar The Wondercat

From Wikipedia, the free encyclopedia

Wikipedia:List of hoaxes on Wikipedia

From Wikipedia, the free encyclopedia

- 1 Episodes: series 1
- 2 Episodes: series 2
- 3 Foreign language version
- 4 References

Episodes: series 1

1. Meet Olimar
2. Bye mum!
3. A long train journey
4. Olimar goes to London

Contents [hide]

- 1 Hoax articles
 - 1.1 Extant for 10+ years
 - 1.2 Extant for 8–9 years
 - 1.3 Extant for 4–7 years
 - 1.4 Extant for 1–3 years
 - 1.5 Extant for less than one year
- 2 Hoax statements in articles
 - 2.1 Extant for 10+ years
 - 2.2 Extant for 8–9 years
 - 2.3 Extant for 4–7 years
 - 2.4 Extant for 1–3 years
 - 2.5 Extant for less than one year
- 3 See also
- 4 References
- 5 Further reading



Hoaxes

Articles that deceptively present false information as a fact.

PROTECTING WIKIPEDIA

Good Editors/Users

Rollbackers

Patrollers

Watchlisters

Readers

Bots/Tools/Blacklists

Cluebot NG

Stiki

ORES

Link-spam blacklist

Page Protection

Account blocking

PROTECTING WIKIPEDIA: *RESEARCH EFFORTS*

Detecting Disinformation

- Vandalism [Adler et al., CICLing 2011] (survey)
- Link-spamming [West et al., OpenSym'11], [West et al., CEAS'11]
- Hoaxes [Kumar, West, Leskovec, WWW'16]

Detecting Deceivers

- **Vandals**: *users who make incoherent and destructive edits* [Kumar, **Spezzano**, Subrahmanian, KDD'15]
- **Spammers**: *users who unsolicitedly promote of some entity* [Green & **Spezzano**, ICWSM'17]
- **Sockpuppets**: *multiple accounts operated by the same user. They are often used to deceive, e.g. to harass other users or to circumvent a block or ban.* [Solorio, Hasan, Mizan, LASM'13], [Yamak et al., WWW'16 Comp.]

Detecting Pages to Protect

- **Page protection**: *placing restrictions on the type of users that can edit a page* [Suyehira & **Spezzano**, CIKM'16], [Suyehira, **Spezzano**, Gundala, SNAM'19]

DETECTING MALICIOUS EDITORS: VANDALS AND SPAMMERS

DETECTING MALICIOUS EDITORS: **VANDALS** AND SPAMMERS

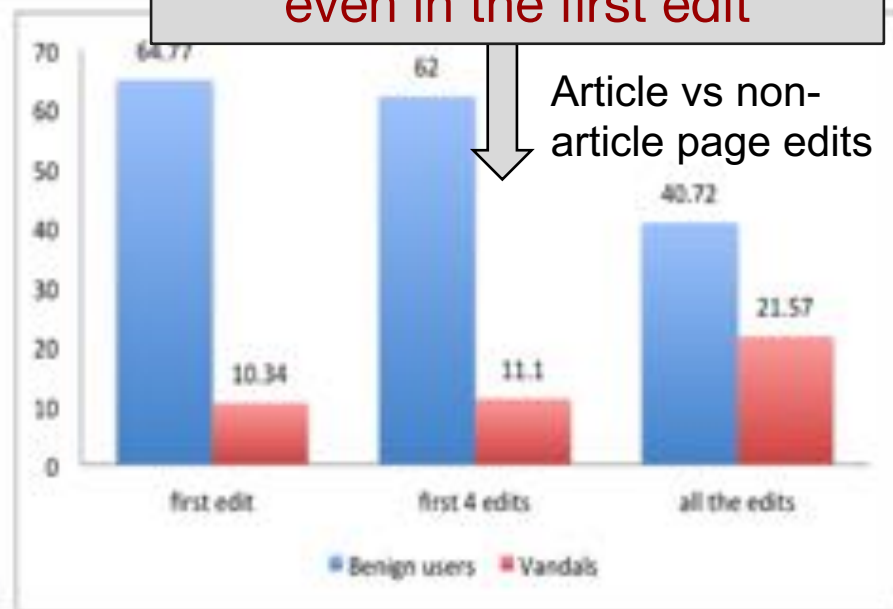
S. Kumar, **F. Spezzano**, and V. Subrahmanian
VEWS: A Wikipedia Vandal Early Warning System
SIGKDD, 2015

HOW DO VANDALS BEHAVE?

Dataset

34,000 Editors Half are vandals
770,000 Edits 160,000 edits by vandals
Jan 2013 - July 2014 new users

Benign users edit more non-article pages than vandals, even in the first edit



Vandals spend less time in editing a new page

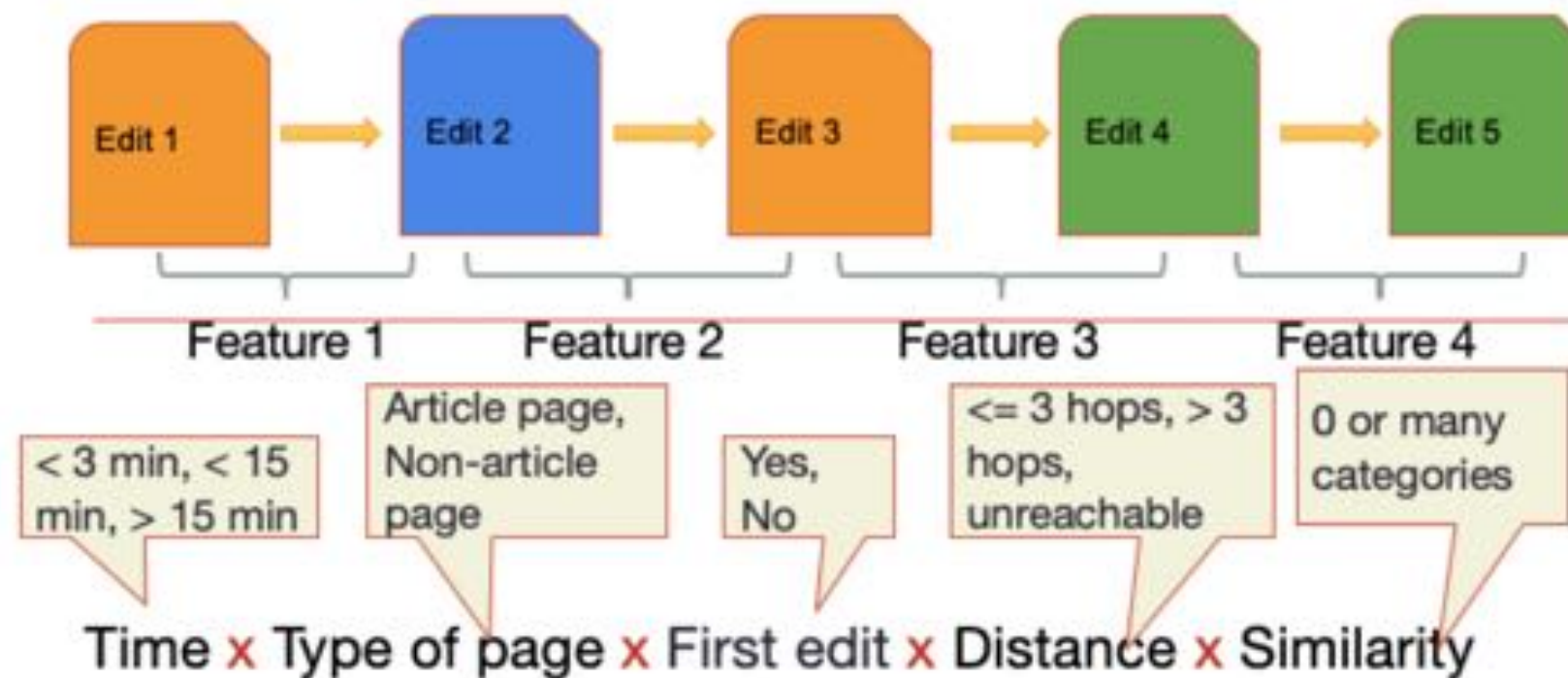


*All results are statistically different with p-values < 0.01

Vandals make faster edits than benign users

VEWS: VANDALS EARLY WARNING SYSTEM

Editor Features



Each edit pair can be in one of 60 categories

Behavioral Features

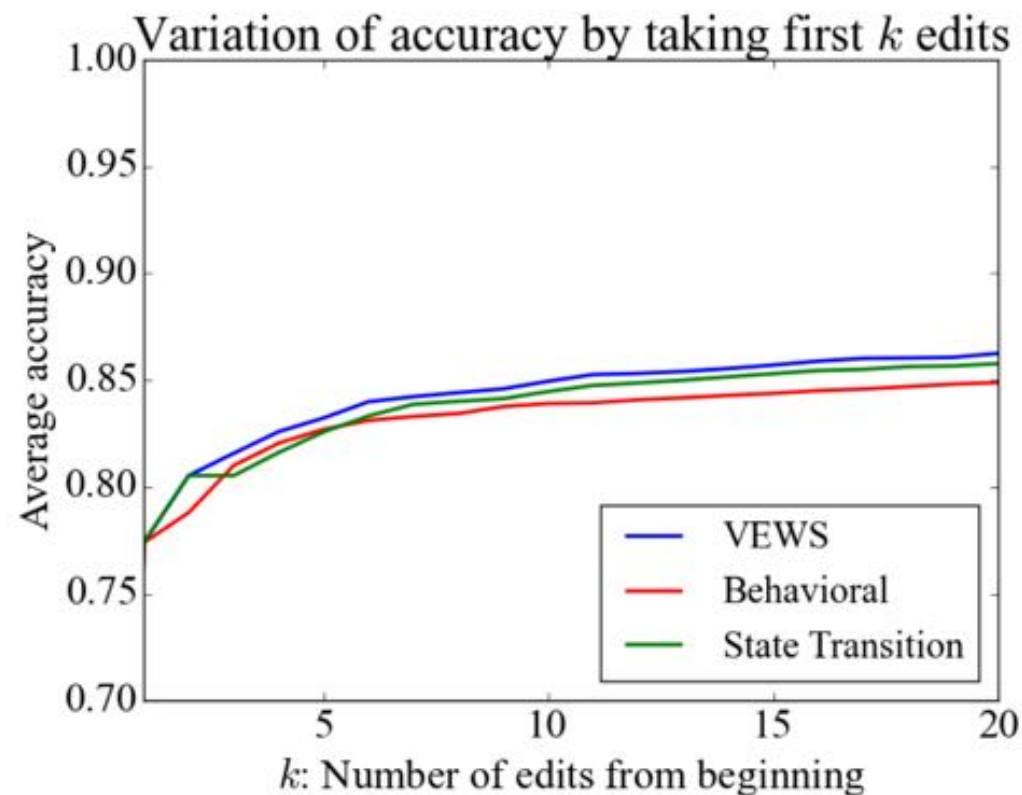
Aggregated features of all benign and all vandal editors

State Transition Features

Individual editor's pattern in sequence of edits

DETECTING VANDALS WITH VEWS

Accuracy of VEWS



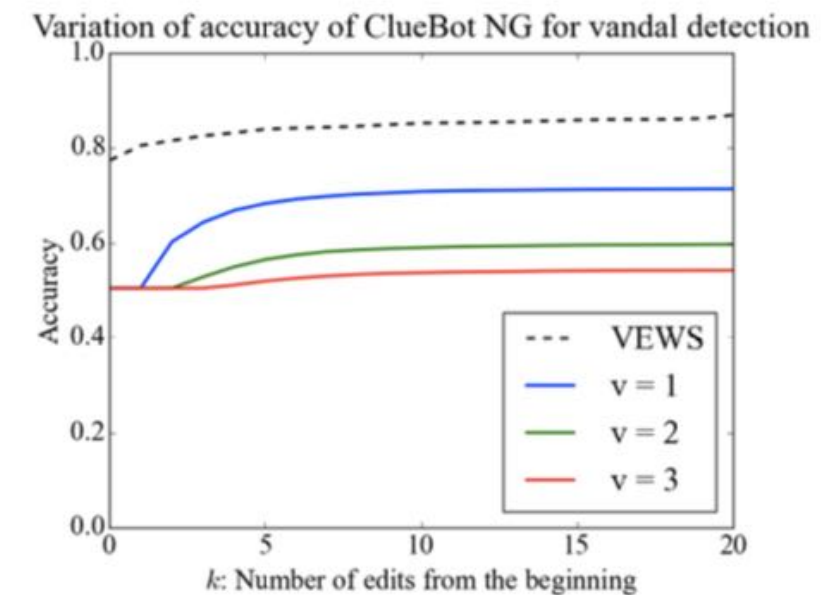
78% accuracy with first edit only.

44% cases vandal identified before first reversion.

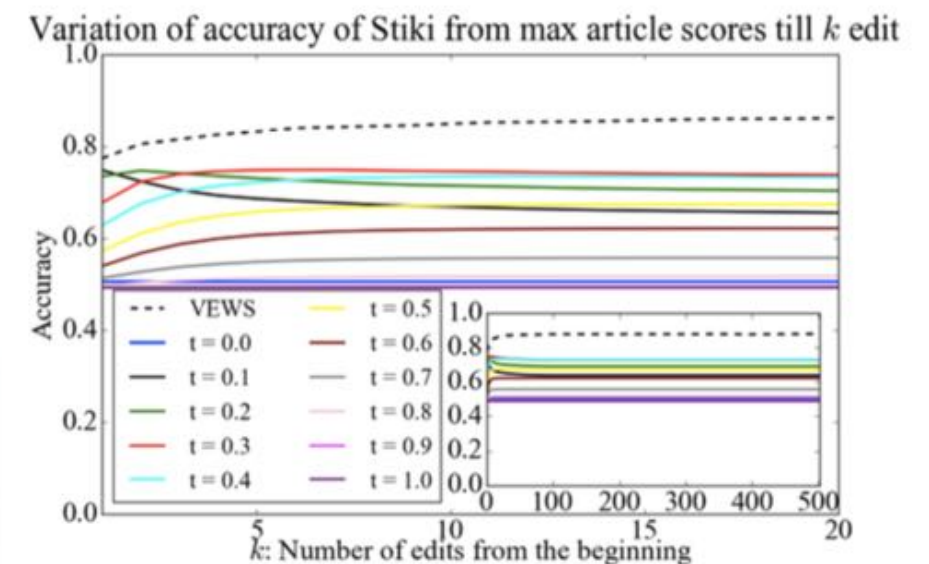
VEWS detects vandals 2.39 edits before ClueBot NG, on avg.

With all edits, VEWS = 88% accuracy

VEWS vs. ClueBot NG



VEWS vs. STiki



DETECTING MALICIOUS EDITORS: VANDALS AND SPAMMERS

T. Green and **F. Spezzano**
Spam Users Identification in Wikipedia via Editing Behavior
ICWSM, 2017

DETECTING WIKIPEDIA SPAMMERS

Spam Dataset

*Our Spam dataset consists of **4.2K** (half spam and half benign) users and **75.6K** edits.*

- All Wikipedia users (up to Nov. 17, 2016) who were blocked for spamming (**2,087 spammers**): “Wikipedians who are indefinitely blocked for spamming” (till Mar 12, 2009); “Wikipedians who are indefinitely blocked for link spamming” (after Mar 12, 2009)*
- An almost equal number of randomly selected benign users (**2,119 benign users**).*
- Up to the last 500 most recent edits for each user.*

Editor Features

Edit size based features

- Average size of edits*
- Standard deviation of edit sizes*
- Variance Significance: $\text{stdDev}/\text{avgSize}$*

Editing time behavior based features

- Average time between edits*
- Standard deviation of time between edits*

Links in edit based features

- Unique link ratio (only link domain)*
- Link ratio in edits*

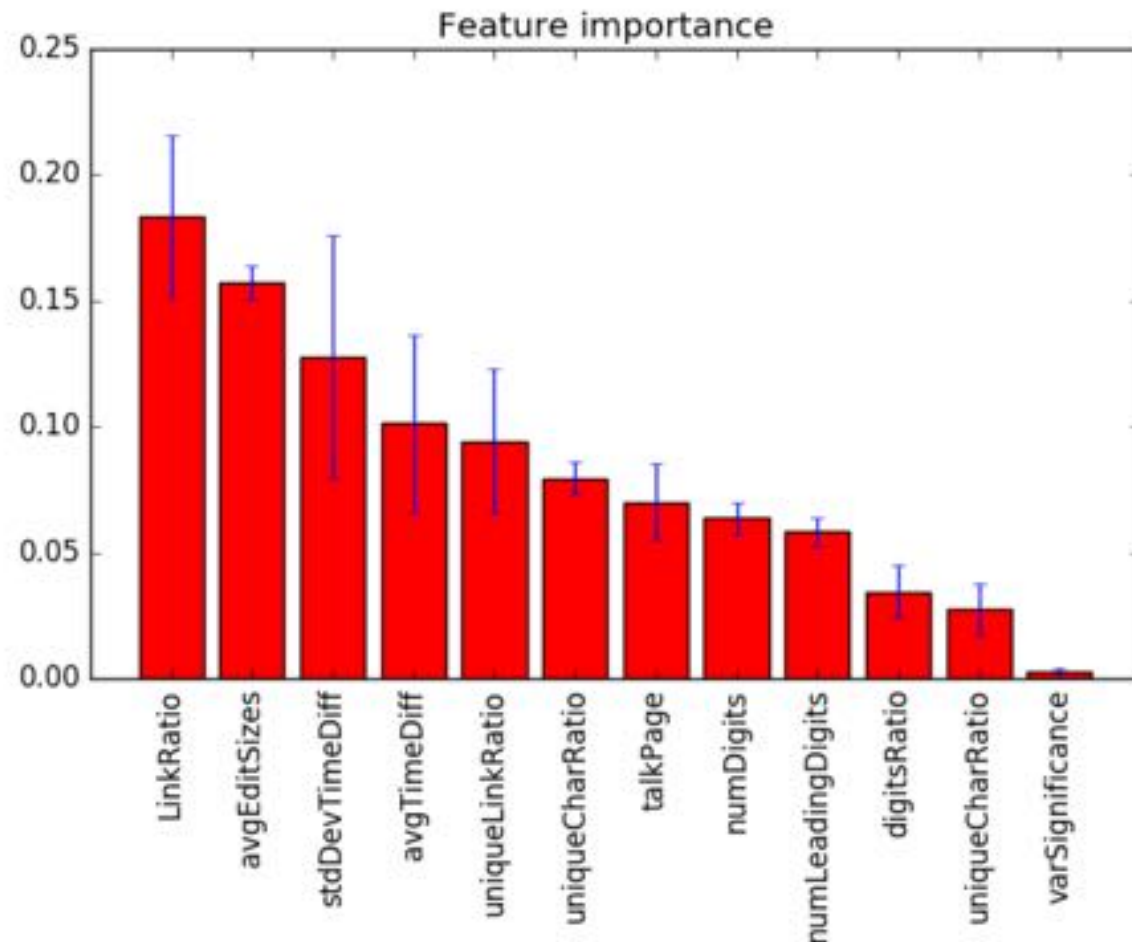
Talk page edit ratio

Ratio of talk pages correspondent to edited pages

Username based features

- Number of digits in a username*
- Ratio of digits in a username*
- Number of leading digits in a username*
- Unique character ratio in a username*

DETECTING WIKIPEDIA SPAMMERS: *FEATURE ANALYSIS*



Top-3 Features

- Link Ratio
- Average size of edits
- Standard deviation of time between edits

Only 30.3% of the users edit talk pages. Among these, the **talk page edit ratio** is higher, on average, for spammers (0.2) than for benign users (0.081).

Username based features contribute to an increase in accuracy prediction by 2.9% (from 77.9% to 80.8%) and mean average precision by 0.019 (from 0.861 to 0.880).

DETECTING WIKIPEDIA SPAMMERS: ACCURACY RESULTS

We can classify spammers from benign users with **80.8%** of **accuracy** and a **mean average precision** of **0.88** on 10-fold cross validation.

Our Features	Accuracy	MAP
SVM	67.0%	0.746
Logistic Regression	79.2%	0.838
K Nearest Neighbor	71.1%	0.733
Random Forest	80.5%	0.856
<u>XGBoost</u>	80.8%	0.880
	Accuracy	MAP
ORES	69.7%	0.695
ORES + Our Features	82.1%	0.886

Table: accuracy and MAP results and comparison with baselines. ORES and ORES + Our Features are computed with XGBoost.

ORES baseline: average and maximum ORES damaging score among all user's edits

Unbalanced Setting: 10% spammers and 90% benign users.

	AUROC
ORES	0.736
Our Features	0.842
ORES + Our Features	0.864

Table: Our features vs. ORES in the unbalanced setting. Everything is computed with XGBoost. The AUROC for the balanced setting is 0.891.

DETECTING PAGES TO PROTECT

K. Suyehira and **F. Spezzano**

DePP: A System for Detecting Pages to Protect in Wikipedia, CIKM, 2016.

F. Spezzano, K. Suyehira, and L. Gundala

Detecting Pages to Protect in Wikipedia across Multiple Languages
In Social Networks Analysis and Mining 9(1): 10:1-10:16, 2019.

WIKIPEDIA PAGE PROTECTION

*When an article is heavily **vandalized**, or because of **libel** or **edit-warring**, administrators may protect the page by restricting its access to "good" users*



The screenshot shows the Wikipedia article for 'Drug'. The page is protected, as indicated by a padlock icon in the top right corner of the article content area. The article title 'Drug' is prominently displayed. Below the title, the text reads: 'From Wikipedia, the free encyclopedia'. A note states: 'For other uses, see [Drug \(disambiguation\)](#).' The main text defines a drug as 'any substance other than food, that when inhaled, injected, smoked, consumed, absorbed via a patch on the skin or dissolved under the tongue causes a physiological change in the body.' It also mentions that in pharmacology, a pharmaceutical drug is a chemical substance used to treat, cure, prevent, or diagnose a disease or to promote well-being. A photograph of a cup of coffee is shown, with a caption below it: 'Caffeine, contained in coffee and other beverages, is the most widely'. The left sidebar contains the Wikipedia logo and various navigation links such as 'Main page', 'Contents', 'Featured content', 'Current events', 'Random article', 'Donate to Wikipedia', 'Wikipedia store', 'Interaction', 'Help', 'About Wikipedia', 'Community portal', and 'Recent changes'. The top navigation bar includes links for 'Not logged in', 'Talk', 'Contributions', 'Create account', and 'Log in'. The article tab is selected, and there are links for 'Read', 'View source', and 'View history'. A search bar is also present.

PROTECTING PAGES: POLICY

There are different levels of page protection:

Fully protected pages can be edited (or moved) only by administrators;

Semi-protected pages can be edited only by *autoconfirmed* users;

Move protection does not allow pages to be moved to a new title, except by an administrator.

Page protections can also be set for different amounts of time, including 24 or 36 hours, or indefinitely.

Currently, all the work is *manually* done by autoconfirmed editors and administrators.

DEPP: DETECTING PAGES TO PROTECT

DePP is the *first* system that is able to decide whether a page should be protected or not in Wikipedia.

Two novel sets of features based on:

1. **Page revision behavior** (features describing how users edit the page)
2. **Page categories** (proxy for page topic)

Advantages: **DePP** does not look at textual content, so it can work with all the different language versions of Wikipedia

DEPP: DETECTING PAGES TO PROTECT

- We build 4 datasets from different Wikipedia versions: English, German, French, Italian.

Dataset	Protected Pages	Non-protected Pages
English	6,799	6,824
German	1722	1706
French	524	512
Italian	171	168

- Each dataset consists of:
 - Half protected and half unprotected Wikipedia pages.
 - All edit protected articles up to Oct. 12, 2016.
 - An almost equal number of randomly selected unprotected pages.
 - Up to the last 500 most recent revisions for each selected page.

DEPP: DETECTING PAGES TO PROTECT

The DePP system achieves at least **0.93 accuracy** across multiple languages.

Accuracy

Dataset	B1	B2	B3	B1+B2+B3	<u>DePP</u>
English	0.56	0.67	0.73	0.80	0.98
German	0.50	-	0.50	0.50	0.98
French	0.50	0.77	0.50	0.77	0.97
Italian	0.50	-	0.50	0.50	0.93

Baselines:

[B1] Number of revisions tagged as "Possible libel or vandalism";

[B2] Number of revisions reverted as possible vandalism by any tool:

Cluebot NG or Stiki (English), Salebot (French), no tool available in German or Italian;

[B3] Number of edit wars between two users in the page: there was an explicit tag for German and Italian Wikipedia.

PAGES PROTECTION AND CONTROVERSIAL TOPICS

Protect from edit-wars → Controversial topics

Detecting controversial pages in Wikipedia by analyzing the page, the editing behavior, or neighborhoods

[Kittur et al., *CHI'07*], [Sepehri Rad and Barbosa, *WikiSym'12*], [Dori-Hacohen and Allan, *ECIR'15*]

Another Baseline: Can we detect pages to protect from their controversy level?

Dataset from [Dori-Hacohen and Allan, *ECIR'15*]:

2060 English Wikipedia pages annotated with their controversy level from 1 to 4 (193 are protected pages).

	AUROC	Avg. Precision
Controversy Level	0.53	0.12

PAGES PROTECTION PAGE POPULARITY

Is there any association between the popularity of a page and page protection?

We defined page popularity as number of page views.

Retrieved from the Wikipedia Clickstream dataset which contains monthly request logs of Wikipedia pages from other Wikipedia pages and any other Web page external to Wikipedia.

	Protected	Unprotected
More than 10 views	21.35%	11.6%
Avg. number of views	71.4K	1.1K

	AUROC	Avg. Precision
Controversy Level	0.53	0.12
Page Popularity	0.60	0.57

CONCLUSIONS

- People trust and read Wikipedia every day. Need to protect Wikipedia from disinformation:
 - a) We presented DePP, an automatic tool to detect pages to protect in Wikipedia across multiple languages.
 - b) We showed that behavior modeling can be very effective to detect malicious editors in Wikipedia (e.g., vandals and spammers).

Drawback of anti-vandal tools [Halfaker et al., 2013]

- Many newcomers face social barriers preventing them from the integration in the editor community, with the consequence of stop editing after a certain period of time.
- **Future work:** Improve our malicious editor detection tools by detecting these users as soon as possible and reduce false positives.

Thank you!



francescaspezzano@boisestate.edu

